UNIVERSITÄT AUGSBURG

Bestimmung intrazyklischer Phasengeschwindigkeiten von Schwimmern im Schwimmkanal mittels vollautomatischer Videoanalyse

D. Zecha, R. Lienhart

Report 2014-04 Juli 2014

INSTITUT FÜR INFORMATIK
D-86135 AUGSBURG

Bestimmung intrazyklischer Phasengeschwindigkeiten von Schwimmern im Schwimmkanal mittels vollautomatischer Videoanalyse

Dan Zecha, Rainer Lienhart
Lehrstuhl für Multimedia und Maschinelles Sehen
Universität Augsburg
{dan.zecha,rainer.lienhart}@informatik.uni-augsburg.de

Einleitung

Für die Leistungsdiagnostik von Schwimmern im Spitzensport bilden Videoaufzeichnungen eine wesentlich Grundlage für die Einschätzung des Bewegungsablaufs. Während im Routinebetrieb der leistungsdiagnostischen Untersuchungen in der Regel nur qualitative Bewertungen der Bewegungsabläufe durchgeführt werden, sind quantitative Auswertungen wegen des enormen personellen Aufwands nur in Einzelfällen möglich. Eine vollautomatische, quantitative Videoanalyse mit dem Ziel, zyklische Strukturen zu erfassen und daraus kinematische Parameter abzuleiten, eröffnet neue Möglichkeiten auf dem Gebiet der Leistungsdiagnostik.
Im Rahmen des BISp-Projekts „Vollautomatische zeitkontinuierliche Bestimmung intrazyklischer Phasengeschwindigkeiten von Schwimmern im Schwimmkanal einschließlich Zugfrequenz und Zuglänge“ [Pro13] wird unter Berücksichtigung jüngster Entwicklungen auf dem Gebiet der Posen- und Bewegungserkennung sowie der Zeitreihenanalyse erforscht, wie sich kinematische Parameter wie Zugfrequenz, Zuglänge und intrazyklische Phasengeschwindigkeiten vollautomatisch, d.h. ohne mühsame und arbeitsintensive manuelle Auswertung, mittels softwarebasierter Auswertung bestimmen lassen.
Alle Parameter werden aus Videodaten bestimmt, die an einem Schwimmkanal aufgenommen werden. Ein Schwimmkanal ist ein kleines Becken in dem eine konstant fließende Strömung erzeugt werden kann. Der Schwimmkanal wird durch eine gläserne Seitenwand sowie von außerhalb des Beckens von Videokameras gefilmt. Ein Sportler führt nun regelmäßige Schwimmbewegungen in einer der vier Schwimmlagen (Brust, Schmetterling, Freistil, Rücken) aus und wird dabei gefilmt. Abbildung 1 zeigt die Perspektive einer Seitenkamera. Bisher dato werden diese Videoaufnahmen von einem Experten (z.B. dem Trainer) ausgewertet. Dieser markiert händisch die Zeitpunkte des Auftretens zuvor festgelegter Schlüsselposen. Anhand der Annotationen können die gewünschten Parameter im Anschluss bestimmt werden. Diese Art der quantitativen Auswertung ist sehr zeitintensive, da der Experte den Großteil der Einzelbilder eines Videos bewerten muss.

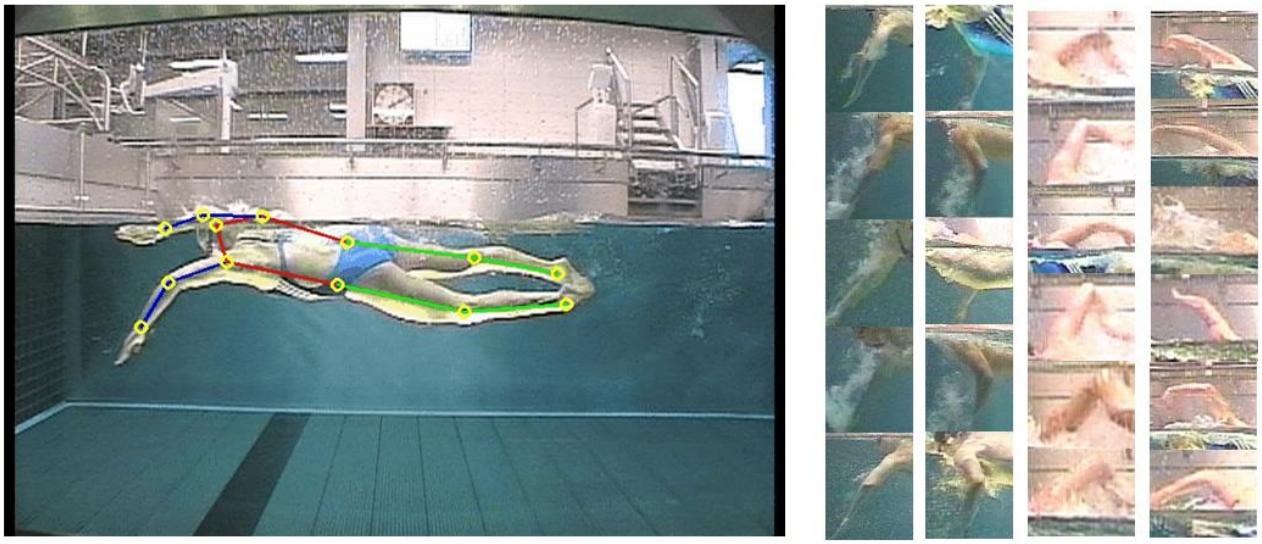

Abbildung 1) Links: Kamerabild einer Schwimmerin im Schwimmkanal (Quelle: IAT Leipzig). Gelbe Kreise markieren die Positionen der Gelenke und damit die Gesamtpose. Rechts: Ausschnitte aus verschiedenen Gruppen von Armlets Trainingsbildern.

Die im Folgenden vorgeschlagene Alternative zur vollautomatischen Bestimmung von Schlüsselposen löst das Problem mittels Algorithmen aus dem Gebiet des Maschinellen Sehens und ist an die Methodik in [Zec12] angelehnt. Das automatisierte Ableiten von Hinweisen zur Verbesserung der Technik eines Schwimmers ist nicht Teil des Systems und fällt in das Gebiet der Trainingswissenschaften.

Um Parameter wie Zugfrequenz und intrazyklische Frequenzen bestimmen zu können, wird das Problem auf die folgenden Betrachtungen reduziert:
Aus einem konstanten Strom von Einzelbildern sollen diejenigen Bilder detektiert werden, die eine vom Experten definierte Pose zeigen. Diese Posen werden als *Schlüsselposen* bezeichnet. Eine Schlüsselpose ist generell anhand von einzelnen Merkmalen der kompletten Pose definiert. Ein solches Merkmal ist zum Beispiel die Position oder der Winkel des Oberarms bei Freistilschwimmern. Die direkte Detektion eines solchen Merkmals mittels eines speziellen Detektors erweist sich in der Regel als schwierig, da es im Kamerabild z.B. zu (Eigen-)Verdeckung des Merkmals durch den Körper des Schwimmers kommt. Zusätzlich erschweren Luftblasen im Wasser (Abbildung 5, Mitte), die vom Schwimmer und den Abschlussgittern im Schwimmkanal erzeugt werden, die direkte Detektion. Es kann somit nicht garantiert werden, dass Schlüsselposen direkt bestimmt werden können. Da eine zyklische Bewegung aber eine vorhersehbare Struktur hat, kann das Auftreten von Schlüsselposen mittels Stützposen innerhalb eines Zyklus vorhergesagt werden. Eine *Stützpose*[1] beschreibt hierbei eine Körperhaltung, die im Gegensatz zu Schlüsselpose einwandfrei innerhalb einer wiederkehrenden Bewegung erkannt werden kann. Das vorgeschlagene

1 Während Schlüsselposen ausschließlich nach Gesichtspunkten der Leistungsdiagnostik und unabhängig von ihrer automatischen Bestimmbarkeit ausgewählt werden, werden Stützposen ausschließlich nach der Leichtigkeit und Güte ihrer automatischen Erkennbarkeit ausgewählt.

Verfahren findet dabei alle Stützposen automatisch, d.h. ohne Zutun eines menschlichen Experten, anhand der Struktur der Trainingsdaten. Im letzten Schritt des Algorithmus wird aus den Zeitpunkten des Erscheinens mehrerer Stützposen eine Vorhersage über das Erscheinen einer Schlüsselpose getroffen.

Allgemeines Vorgehen. Zunächst wird ein Modell zur Detektion eines Schwimmers im Schwimmkanal trainiert. Das Modell wird mit weiteren Detektoren für die Arme des Schwimmers erweitert. Hierzu werden alle möglichen Armhaltungen aus einer Datenbank mittels Clusteranalyse in Gruppen eingeteilt. Eine speziell dafür entworfene Zielfunktion garantiert dabei, dass nur äußerlich und (im Zyklus eines Schwimmzugs) zeitlich ähnliche Armkonfigurationen in einer Gruppe zusammengefasst werden. Jeder Armdetektor wird relativ zum Körperdetektor ausgewertet und funktioniert damit wie ein Sensor für die An- oder Abwesenheit eines Arms. Der Detektor erzeugt einen hohen Ausgabewert, wenn der Schwimmer sich in der entsprechenden Körperhaltung befindet. Fährt er mit der Bewegung fort, so verändert sich auch die Armhaltung und der Ausgabewert des Armdetektors sinkt. Durch die Beobachtung der Ausgabewerte für jeden Armdetektor kann über die Zeit bestimmt werden, wann eine Stützpose am wahrscheinlichsten auftritt. In einem letzten Schritt wird mittels eines statistischen Schätzverfahrens der Zeitpunkt des Auftretens einer Schlüsselpose gefolgert.

Glossar. Innerhalb einer regelmäßigen zyklischen Bewegung ist ein *Zyklus/Zug* die kleinste Einheit. Ein Zyklus ist definiert als die Zeit, die zwischen dem Auftreten einer Schlüsselpose und ihrem frühesten Wiederauftreten vergeht. Alle vier Grundschwimmarten bestehen aus regelmäßigen Zyklen/Zügen. Dabei können Brust und Schmetterling als *symmetrischer* und Freistil und Rücken als *antisymmetrischer* Schwimmstil bezeichnet werden. Der wesentliche Unterschied besteht in der Ausführung: Während beide Körperhälften in den ersten beiden Lagen zu jedem Zeitpunkt gespiegelt dieselbe Bewegung ausführen (symmetrisch zur Längsachse des Schwimmers) und daher immer die gleiche Schlüsselpose zeigen, wechseln sich die Körperhälften in den Lagen Rücken und Freistil wechselseitig ab; während eine Körperhälfte sich unter Wasser in der Druckphase befinden, ist die andere Seite in der Ruhephase über Wasser. Die Schlüsselpose tritt also etwa jeden Halbzyklus einmal auf, und zwar einmal auf der rechten und einmal auf der linken Körperseite. Die explizite Unterscheidung ist deshalb wichtig, weil die meisten (gradienten-basierten) Detektoren Probleme haben, beide Körperseiten voneinander zu unterscheiden, wenn der Schwimmer von der Seite gefilmt wird. Dieser Artikel beschäftigt sich im Folgenden nur mit Freistilschwimmern, da die Detektion von Schlüsselposen in den antisymmetrischen Schwimmstilen etwas komplizierter, aber auch allgemeiner ist als für symmetrische Schwimmstile.
Die Körperhaltung eines Menschen wir im Allgemeinen als *Pose* bezeichnet. Im Kontext von Bildverarbeitung und maschinellem Sehen ist die Pose einer Person definiert

als die Position seiner Gelenke in einem Bild. In Abbildung 1 ist die Pose der Schwimmerin mit gelben Kreisen markiert. Gruppen von Gelenkannotationen werden als *Konfigurationen* bezeichnet. Dies schließt die Gesamtkonfiguration des Schwimmers (sein gesamtes Skelett) genauso mit ein wie etwaige Untergruppen, z.B. Konfigurationen aus Schultergelenk, Ellenbogen und Handgelenk zur Darstellung eines Arms.

Detektion von Schlüsselposen

Die folgenden Abschnitte beschreiben das Vorgehen zur Detektion von Schlüsselposen. Neben den Modellen selbst wird auch die Trainingsdatenbank sowie Merkmale und der Prozess der Modellbildung und Anwendung kurz beschrieben.

Datenbank. Jedes gelernte Modell im beschriebenen Verfahren wird Datengetrieben erzeugt. Anstatt also das Aussehen eines Schwimmers direkt mittels eines deskriptiven Modells zu beschreiben, wird ein Detektor mittels eines Algorithmus auf Basis einer Datenbank von Trainingsbildern erlernt. Hierzu wurden die Gelenkkonfigurationen in 1200 Bilder von Freistilschwimmern (3 männliche, 5 weibliche) von einem Experten annotiert. Die Bilder zeigen insgesamt 20 Armzüge, haben eine Auflösung von 720x576 Pixeln und wurden aus einer Seitenansicht auf den Schwimmkanal mit 25Hz (interlaced, insgesamt also 50 Halbbilder) aufgenommen. Es wurde versucht, die meisten offensichtlichen Variablen abzudecken, die Bildqualität und Konfigurationsraum beeinflussen. Dazu gehören Bilder von Schwimmern verschiedenen Geschlechts mit unterschiedlichem Körperbau sowie unterschiedlicher Beleuchtung des Kanals und Fließgeschwindigkeiten des Wassers (zwischen 1m/s und 1.75 m/s). Aus allen Bildern wurden die Konfigurationen für Kopf und Oberkörper einschließlich der Knie für den Körperdetektor sowie die Armkonfigurationen (Schulter, Ellenbogen, Handgelenk) für die Clusteranalyse verwendet. Abbildung 1 zeigt Beispiele aus verschiedenen Gruppen.
Eine Besonderheit bei der Annotation von antisymmetrischen Posen ist die Verdeckung von einzelnen Gelenken z.B. durch den Schwimmer selbst oder durch Lichtbrechung an der Wasserkante, die zum Teil komplette Extremitäten verdecken kann. Der menschliche Experte wird in diesem Fall die Position eines Gelenkes durch Beobachtung der zeitlich benachbarten Bilder interpolieren. Diese Art der Annotation führt zu inakkurateren Annotationen, denen gegenüber das Training der Modelle robust sein sollte.

Merkmale, lineare Filter und ihre Anwendung. Die Bildung eines Modells zur Objekterkennung auf Basis von reinen Pixelvergleichen ist auf dem Gebiet des maschinellen Sehens für gewöhnlich nicht praktikabel, da ein Pixelvergleich relevantere visuelle Informationen über das Objekt nicht von irrelevanteren Informationen unterscheidet und zudem Invarianz gegenüber affinen Transformationen des Objekts sowie Ausleuchtung nicht gewährleistet ist.

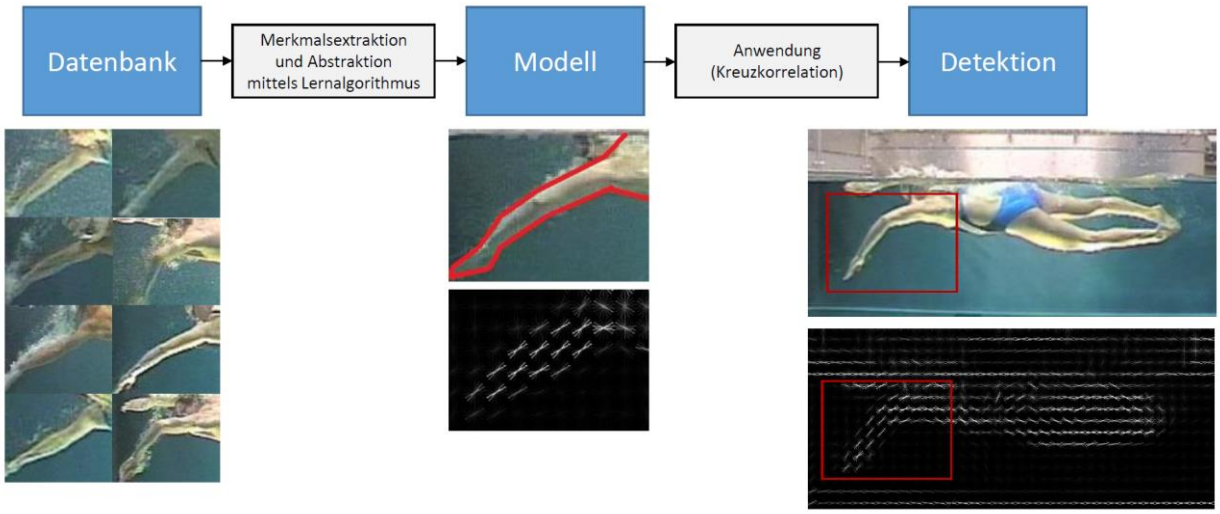

Abbildung 2) Überblick über das vorgeschlagene System: Aus einer Datenbank von Beispielbildern (hier: Armkonfigurationen, links) werden Merkmalskarten extrahiert und es wird ein Modell trainiert (Mitte, unten). Die Struktur eines Arms der entsprechenden Konfiguration (mit roter Umrandung verdeutlicht) ist im Detektor klar erkennbar. Das Modell wird mittels Kreuzkorrelation mit einem Bild (rechts, oben) im Merkmalsraum (rechts, unten) verglichen. Der größte Korrelationswert ergibt dann die Detektion des Arms (rotes Rechteck, rechts).

Eine Lösung dieses Problems bietet die Überführung des Bildes in einen Merkmalsraum. Ein Merkmalsraum bietet eine Abstraktion von den reinen Pixelwerten eines Objekts mit der Aufgabe, Gemeinsamkeiten zu betonen und Unterschiede zu unterdrücken. Anschaulich gesprochen sollte eine guter Merkmalsraum für Schwimmer die Möglichkeit bieten, die Silhouette und speziell die Arme besonders gut zu detektieren und dabei möglichst die Farbe der Schwimmbekleidung zu vernachlässigen, damit das Modell allgemein auf männlichen wie auch auf weiblichen Schwimmern mit verschiedenfarbigen Badehosen und Anzügen gleich gut funktioniert.

In den letzten Jahren haben sich auf dem Gebiet der Objekterkennung Gradientenhistogramme als gutes Merkmal herauskristallisiert (sog. HoG-Merkmale [Dal05]). Ein einzelner Gradient beschreibt dabei die Kontrastunterschiede innerhalb einer sehr kleinen Bildregion. Starke Kanten erzeugen starke Gradienten, während Gradienten in homogeneren Bildbereichen wesentlich schwächer sind. Innerhalb von rechteckigen Regionen der Größe 8x8 Pixel können alle Gradienten zu einem Histogramm über diskretisierte Gradientenrichtungen zusammengefasst werden. Mittels geeigneter Normalisierung über mehrere benachbarte Gradientenhistogramme entsteht eine (gegen kleine Veränderungen robuste) Merkmalskarte als Repräsentation eines Bildausschnitts. Gradientenbasierte Merkmalskarten haben die Eigenschaft, bis zu einem gewissen Grad gegen affine Transformationen des zugrunde liegenden Objekts sowie gegen Änderungen in der Beleuchtung invariant zu sein.
Abbildung 2 visualisiert eine solche Merkmalskarte für einen Arm und einen Schwimmer im Schwimmkanal. Da diese HoG-Merkmale im Wesentlichen Kanteninformationen bzw. Kontrastunterschiede in einem Bild darstellen, ist die Silhouette des Schwimmers/Arms in beiden Merkmalskarten deutlich erkennbar, während die Gradienten in einer homogeneren Region wie dem Wasser kaum vorhanden sind.

Mittels einer Menge von Merkmalskarten, die aus ähnlichen Bildausschnitten berechnet wurden (z.B. Arme von verschiedenen Schwimmern in der gleichen Pose, siehe Abbildung 2 links) kann weiter abstrahiert werden, um einen finalen Armdetektor für die die Arme einer bestimmten Pose zu erhalten. Dies wird mit einem maschinellen Lernalgorithmus wie einer linearen Support Vector Machine (SVM, [Wap95]) erreicht. Die Eingabe in eine SVM sind im beschriebenen Beispiel verschiedene Merkmalskarten, die aus sich ähnlichen Bildausschnitten generiert wurden. Das Ergebnis es Algorithmus ist eine einzige Merkmalskarte (auch als linearer Filter bezeichnet), die eine Abstraktion aller Trainingsbeispiele darstellt. Mittels eines solchen linearen Filters lässt sich ein Objekt (wie z.B. der Schwimmer oder der Arm eines Schwimmers) in einem beliebigen Testbild detektieren. Dabei muss das Testbild allerdings erst in den gleichen Merkmalsraum wie der lineare Filter überführt werden, da ein Vergleich zwischen beiden nur in diesem möglich ist.
Ein linearer Filter auf Basis von Gradientenhistogrammen wir mittels mathematischer Kreuzkorrelation mit einem Testbild verglichen. Dabei wird jeder Bildausschnitt im Testbild (mit derselben Größe wie der Filter) mit dem Filter verrechnet. Das Resultat ist ein Korrelationswert für jeden Ausschnitt. Je ähnlicher der Bildausschnitt mit dem im linearen Filter gelernten Bildinhalt ist (beides wird im Merkmalsraum verglichen), desto höher ist auch der Korrelationswert. Je unähnlicher sich beide sind, desto kleiner wird der Wert. Abbildung 2 veranschaulicht diesen Prozess. Der lineare Filter für den Arm wird an der Stelle des roten Quadrats den höchsten Korrelationswert erzeugen.
Erwähnenswert ist an dieser Stelle, dass ein linearer Filter nicht immer funktionieren muss. Ein gutes Resultat hängt dabei von vielen Faktoren ab. Neben einer gut gewählten Trainingsmenge für den Filter und der richtigen Wahl des Trainingsalgorithmus wirkt sich auch die Qualität des Testbildes aus (starke Rauschen durch Luftblasen im Wasser, tatsächliche Pose des Schwimmers, Verdeckung von Teilen). Auch zeigt sich, dass bestimmte lineare Filter besonders gut für langsame Schwimmer funktionieren, nicht aber für schnellere, und umgekehrt.
Auf dem Gebiet der menschlichen Posenschätzung in Bildern haben sich in den letzten Jahren spezielle Begriffe für lineare Filter auf Basis von Gradientenhistogrammen herauskristallisiert. So beschreibt der Begriff *Poselet* [Bou09] immer einen linearen Filter, der für eine beliebige Gelenkkonfiguration eines Menschen in einer spezifischen Pose trainiert worden ist. Demnach ist ein *Armlet* [Gki13] ein Filter zur Detektion von menschlichen Armkonfigurationen. Im Folgenden werden die Begriffe Poselet, Armlet und (linearer) Filter synonym verwendet.

Schlüsselposendetektion. Angelehnt an [Fis73, Fel10] besteht ein kompletter Detektor für einen Freistilschwimmer aus insgesamt 16 Poselets. Ein Wurzelposelet wird für die Konfiguration eines kompletten Athleten trainiert und liefert eine erste Abschätzung für die Position des Schwimmers. Weitere 15 Armlets komplementieren das Modell. Sie werden aus der Menge aller aus den Trainingsbildern ausgeschnittenen Armkonfigurationen für rechte und linke Arme trainiert. Alle Bildausschnitte

werden mittels einer Clusteranalyse (k-means-Algorithmus) in Gruppen aufgeteilt. Dabei sichert eine geeignete Zielfunktion, dass alle Ausschnitte innerhalb einer Gruppe ähnliche Armposen zeigen. Abbildung 2 zeigt Beispiele aus Trainingsmengen für vier verschiedene Armlets.
Jede Gruppe von Armbildern definiert im finalen Modell eine Stützpose. Es wird deutlich, dass Stützposen damit nicht von einem menschlichen Experten definiert sind, sondern von einem Algorithmus „gewählt“ werden. Aus jeder Gruppe wird anschließend ein Armlet trainiert. Da die Trainingsdaten nur Bilder aus der Seitenperspektive enthalten, sind in den Gruppen Armkonfigurationen von rechten und linken Armen vermischt. Die final trainierten Armlets können damit zum Testzeitpunkt nicht zwischen rechter und linker Körperseite unterscheiden.
Während der Körperdetektor zum Testzeitpunkt mit dem kompletten Bild korreliert werden muss, um den Schwimmer zuverlässig zu detektieren, werden die Armlets nur in beschränkten Bildbereichen ausgewertet, die mit dem detektieren Körper konform sind. Die Grundidee hierbei ist es, nicht nach einem Arm in Bildbereichen zu suchen, in denen sich kein Schwimmer befindet. Der Bereich für die Auswertung eines Armlets wird bei der Extraktion der Bildausschnitte relativ zur Position des Schwimmers im Training gelernt. Zum Testzeitpunkt wird ein Armlet nur in dem Bereich ausgewertet, in dem der Arm vermutet wird. Damit funktionieren alle Armlets wie ein Netz aus Sensoren: Wenn eine für ein Armlet trainierte Konfiguration in einem Video tatsächlich auftritt, so ist der maximale Korrelationswert zwischen Armlet und Bild (innerhalb des erlaubten Auswertungsbereichs) hoch. Fährt der Schwimmer mit seiner Bewegung fort und verändert seine Gesamtpose, so sinkt der Korrelationswert wieder.
Um zu entscheiden, wann eine Stützpose tatsächlich auftritt, werden pro Armlet alle maximalen Korrelationswerte (einer pro Bild und detektiertem Schwimmer) in einer Zeitreihe (– auch Posensignal genannt –) aggregiert. Um hochfrequentes Rauschen aus jedem Posensignal zu entfernen, werden diese mittels eines Gaußfilters geglättet. Der Zeitpunkt des Auftretens einer Stützpose ist dann durch jedes lokale Maximum im Posensignal gegeben, also den Zeitpunkten, an dem der Korrelationswert am größten ist. Abbildung 3 veranschaulicht diesen Prozess der Stützposengenerierung für ein Armlet. Für anti-symmetrische Schwimmstile erzeugt das Modell pro Zyklus jeweils zwei Stützposen (je eine für die linke und rechte Körperhälfte), für symmetrische Schwimmstile tritt eine Stützpose nur einmal pro Zyklus auf. Für gewöhnlich enthalten vor allem schlechtere Posensignale viele falsche Stützposen. Die Liste der Stützposen wird daher mittels einer einfachen Heuristik gesäubert, die alle Stützposen entfernen, die die Zugfrequenz des Schwimmers über ein normales Maß hinaus zu stark ansteigen lassen. Nach der Säuberung der Signale hat das Modell insgesamt 15 verschiedene Mengen von Stützposen berechnet.

Um von einer Sammlung von Stützposen das Auftreten einer Schlüsselpose zu schätzen, wird zuerst der Schwimmerdetektor auf 30 Testvideos angewendet und

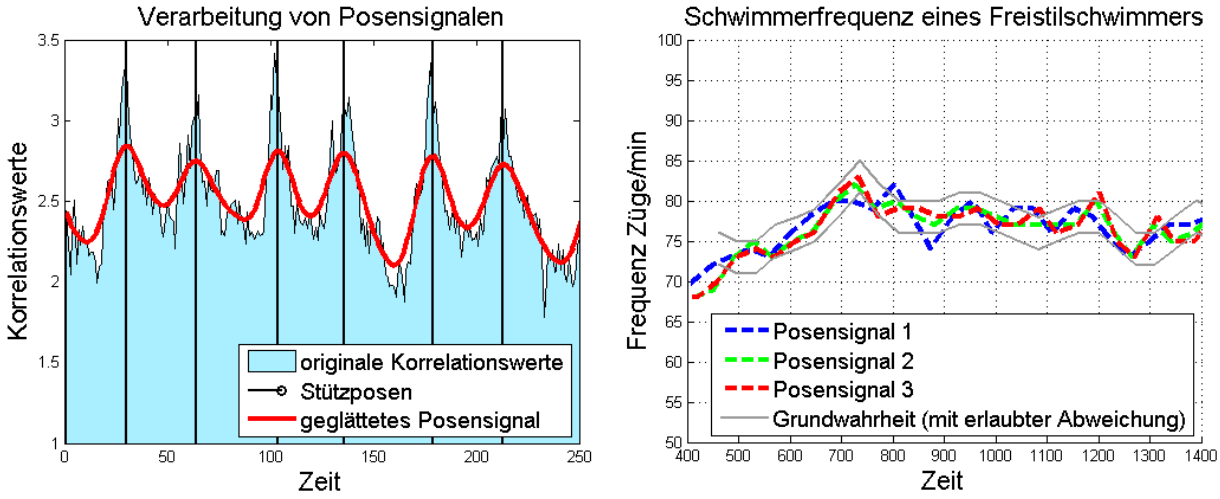

Abbildung 3) Links: Generierung von Stützposen (schwarz) aus geglätteten (rot) Korrelationswerten eines Armlets mit einem Testvideo (blau). Rechts: Berechnung der Schwimmerfrequenz aus den besten 3 Posensignalen. Der von grauen Graphen eingefasste Korridor stellt die erlaubte Abweichung von der Grundwahrheit dar.

die Stützposen nach dem oben beschriebenen Verfahren ermittelt. Da nicht alle Armlets immer gleich zuverlässig funktionieren, wird ein Gütekriterium auf Basis der Selbstähnlichkeit eines Posensignals definiert und die Posensignale gemäß ihrer Güte sortiert. Während bessere Posensignale sehr regelmäßige und korrekte Stützposen liefern, können schlechte Posensignale schnell Falschdetektionen erzeugen. Die Auswirkung von guten und schlechten Posensignalen wird in der Evaluation genauer untersucht. Für alle Testvideos annotiert ein menschlicher Experte zudem die Bildernummer, in denen eine Schlüsselpose auftritt. Gegeben der annotierten Grundwahrheiten des Experten wird für alle Schlüsselposen eines Armlets die Parameter einer (Normal-) Verteilung geschätzt. Der Mittelwert dieser Verteilung gibt an, wie weit eine Schlüsselpose innerhalb eines Zyklus von einer Stützpose entfernt ist, während die Standardabweichung einen Konfidenzbereich um den Mittelwert definiert. Sowohl Mittelwert als auch Standardabweichung werden dabei mit der entsprechenden Zuglänge normalisiert, sodass die Abstände immer unabhängig von der Zugfrequenz eines spezifischen Schwimmers geschätzt werden. Pro Armlet bilden Mittelwert und Konfidenz zusammen einen Schlüsselposenschätzer; insgesamt entstehen also 15 solcher Einzelschätzer.

Die finale Schlüsselpose wird nun mittels verschiedener statistischer Schätzverfahren ermittelt:

- Die maximum-likelihood Methode (ML Hypothese) schätzt die Position einer Schlüsselpose innerhalb eines Testvideos nur anhand der ermittelten Schlüsselposenschätzer. Jedes an der Gesamtschätzung beteiligte Armlet gibt dabei eine eigene Vorhersage für das Auftreten einer Schlüsselpose ab. Die finale Position wird dann aus allen Teilschätzungen gemittelt.
- Die a-priori-Verteilung (prior Hypothese) nimmt an, dass der Experte für ein Testvideo ein Vorkommen der entsprechenden Schlüsselpose markiert hat und schätzt aufgrund dieser Basisannotation und den Stützposen dann jedes weitere Vorkommen der Schlüsselpose. Die trainierten Schlüsselposenschätzer werden in dieser Vorhersage nicht berücksichtigt.

- Die Maximum-a-posteriori-Methode (MAP Hypothese) vereinigt maximum-likelihood und a-priori Methode. Zusätzlich zu den einzelnen Schlüsselposen-schätzern wird also auch eine einzige Expertenannotation berücksichtigt.

Alle Schätzmethoden werden im folgenden experimentellen Abschnitt ausgewertet.

Evaluation

Die Leistungsfähigkeit der vorgeschlagenen Modelle wird anhand von 30 Videos (720x576@50i) evaluiert und diskutiert, die die linke Körperseite verschiedener Schwimmer (6 männliche, 8 weibliche, Alter zwischen 15 und 25, unterschiedlicher Körperbau) im Schwimmkanal bei langsam ansteigender Fließgeschwindigkeit zwischen 1m/s und 1.75m/s zeigen. Ein menschlicher Experte hat zum Zweck der Evaluation alle Bilder, die eine bestimmte, exemplarisch ausgewählte Schlüsselpose zeigen (hier: Vertikaldurchgang des Oberarms unter Wasser, insgesamt 424 Annotationen), händisch annotiert. Abbildung 5 zeigt diese Pose für zwei verschiedene Schwimmer und für beide Körperhälften.
Als Schwimmerdetektor wurde ein 16-teiliges Poselet Modell trainiert. Das Wurzelposelet detektiert den kompletten Schwimmer, während die restlichen Poselets verschiedene Armkonfigurationen aus verschiedenen Zeitintervallen eines Zyklus abdecken und daher auch als Armlets bezeichnet werden. Die Effizienz dieser 15 Armlets wird im Folgenden indirekt über ihre Detektionsleistung von Schlüsselposen analysiert.
Die Schätzung der Schlüsselposen wird mittels eines Kreuzvalidierungsverfahrens evaluiert, bei dem jeweils 29 Videos zur Erlernen der Schlüsselposenschätzer verwendet werden, welcher dann auf dem verbliebenen Video getestet wird. Sofern nicht anders erwähnt sind, alle Ergebnisse stets über alle Testvideos gemittelt. Tabelle 1 fast alle Testergebnisse zusammen.

Performanzmaße. Generell muss zwischen verschiedenen Typen von Detektionen unterschieden werden: Falls die Schätzung für das Auftreten einer Schlüsselpose weniger als eine gegebene Zahl *x* an Halbbilder von einer annotierten Grundwahrheit abweicht, wird die Schätzung als korrekt gewertet (richtig positiv, TP). Liegt die Schätzung außerhalb von *x* Halbbildern um eine Grundwahrheit, gilt sie als Falschdetektion (falsch positiv, FP). Ein Grundwahrheit, die keiner Schätzung zugeordnet werden kann, ist falsch negativ (FN). Im Folgenden wird die mittels der Zuglänge eines Schwimmers normalisierte Abweichung einer Schätzung relativ zu ihrer Grundwahrheit (d.h. Abweichung von der Grundwahrheit in Prozent der aktuellen Zuglänge, angetragen auf der x-Achse) mit dem Recall des Systems (y-Achse) verglichen. Der Recall ist definiert als recall=TP/(TP+FN) und ist ein Indikator für die Anzahl korrekt geschätzter Schlüsselposen. Eine Recall von 1 bedeutet, dass alle Schlüsselposen richtig geschätzt wurden; er fällt mit

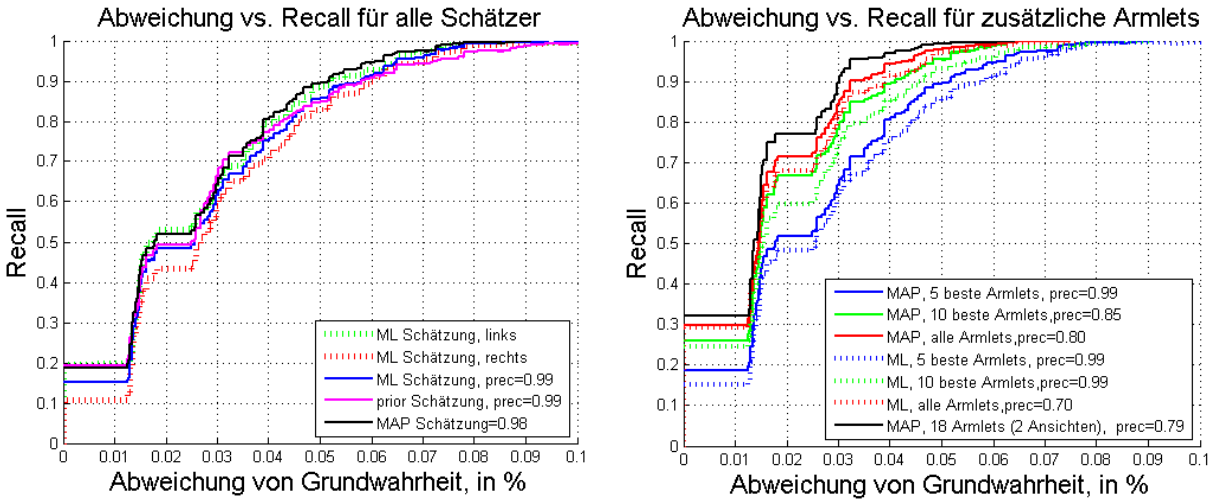

Abbildung 4) Detektionsleistung verschiedener Schätzer relativ zur Abweichung von der Grundwahrheit.

jeder nicht erkannten Schlüsselpose. Da auch der Mensch beim annotieren Fehler machen kann, wird eine gewisse Abweichung von der Grundwahrheit erlaubt. Daher wird jede Abweichungs-Recall Kurve bei einer Abweichung von 3% (x=0.03) ausgewertet. Diese Abweichung entspricht einer Falschannotation von ±1 bis ±2 Halbbildern, was im Bereich des menschlichen Fehlers liegt. Für jeden Graphen wird außerdem die Precision des Schätzers angegeben. Sie ist definiert als precision=TP/(TP+FP) und ist ein Indikator für die Anzahl der Falschdetektionen des Systems. Eine Precision von 1 entspricht keiner Fehldetektion; kleinere Precision-Werte werden durch Falschdetektionen hervorgerufen.

Auswertung von ML, prior und MAP Hypothese. Zuerst wird die Detektionsleistung des maximum-likelihood Schätzers bewertet. Abbildung 4 visualisiert die Abweichung der Schätzung relativ zur Grundwahrheit. Mittels der besten fünf Stützposen wird ein Recall von 0.61 erreicht (, d.h. 61% der Schlüsselposen wurden richtig erkannt). Die Precision von 0.99 impliziert eine sehr kleine Anzahl von Falschdetektionen. Diese entstehen, weil die fünf besten Posensignale nicht fehlerfrei sind.
Abbildung 4 trennt die ML Schätzung (blau) für beide Arme separat auf (grüner Graph = linke Körperhälfte, roter Graph = rechte Körperhälfte). Es zeigt sich, dass Schlüsselposen auf der der Kamera zugewandte Körperhälfte geringfügig besser erkannt werden. Dies ist der Tatsache geschuldet, dass ein menschlicher Experte die Schlüsselposen auf der rechten Körperseite aufgrund von Verdeckung nicht immer exakt sehen kann und den Zeitpunkt ihres Auftretens oft selbst abschätzen muss. Es kann daher angenommen werden, dass der schlechtere Recall auf der rechten Körperseite dem größeren Annotationsfehler des menschlichen Experten geschuldet ist.
In Abbildung 4 wird zudem die Leistung der Prior Hypothese sowie des gesamten MAP Schätzers ausgewertet. Während die a-priori Schätzung wenig überraschend etwas besser ist als die allgemeinere ML Schätzung, werden beide hinsichtlich der Detektionsleistung von der Gesamthypothese etwas übertroffen.

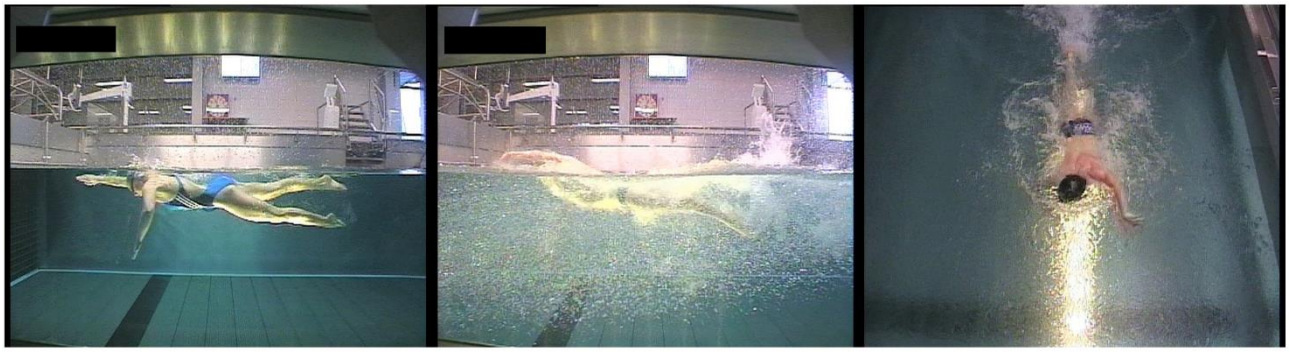

Abbildung 5) Links+Mitte: Schlüsselpose „Vertikaldurchgang Oberarm unter Wasser“ auf beiden Körperseiten. Die Bilder zeigen zwei verschiedene Schwimmer in unterschiedlichen Fließgeschwindigkeiten des Kanals. Rechts: Zusätzliche Kameraperspektive. (Quelle: IAT Leipzig)

Verbesserung des Recalls. Um die Detektionsleistung des Systems zu verbessern, wurden sukzessiv weitere „schlechte“ Stützposenschätzungen (aus schlechten Posensignalen) zur MAP Schätzung hinzugefügt. Während damit der Recall auf 85% angehoben werden kann, fällt die Precision als Nebeneffekt auf 80%. Dies ist wenig überraschend: Während auch schlechte Posensignale in korrekten Abschnitten gute Schätzungen für Stützposen liefern, die wiederrum die Vorhersagen für Schlüsselposen verbessern, führen sie auch eine größere Anzahl schlechter Stützposen in die Gesamtabschätzung ein, was zu einer größeren Anzahl an Falschdetektionen führt. Überraschenderweise hält der ML Schätzer die Precision etwas länger auf einem hohen Niveau (die besten 10 Posensignale erzeugen immer noch eine Precision von 0.99), auch wenn er generell etwas schlechtere Ergebnisse erzeugt als die MAP Hypothese (siehe grüne Graphen in Abbildung 4, rechts).

Zusätzliche Kameraansicht. Die Erkenntnis über die Möglichkeit der Verbesserung von der Schätzung von Schlüsselposen mittels zusätzlicher (wenn auch schlechter) Posensignale inspiriert folgendes Experiment. Ein zusätzliches 7-teiliges Poseletmodell (ein Wurzelposelet für den gesamten Schwimmer plus 6 Armlets) wird für eine zweite Kameraperspektive auf Schwimmer im Schwimmkanal trainiert (siehe Abbildung 4, rechts). Während diese Kamera aufgrund von starker Wasserbewegung keine Posen unter Wasser zeigt, ist sie zur Detektion von Armkonfigurationen über Wasser sehr gut geeignet. Das trainierte Poselet-Modell verhält sich für diese Perspektive trotz anti-symmetrischem Schwimmstil wie ein Modell für einen symmetrischen Schwimmstil, da jeder Arm von einer eigenen Gruppe von Armlets detektiert wird. Da aber jeweils ein Paar von Armlets nur vertikal gespiegelte Versionen voneinander sind, können die Posensignale dieser Paare zusammengefasst werden um eine anti-symmetrischen Output zu simulieren. Die dadurch entstehenden 3 Posensignale werden zusätzlich mit den 15 Signalen des ursprünglichen Modells ausgewertet. Als Ergebnis steigt der Recall der gesamten MAP Abschätzung für die Schlüsselposen bei einer Abweichung von 3% um zusätzliche 4.5% auf 0.89

	ML5	ML5_le	ML5_ri	ML10	ML15	MAP5	MAP10	MAP15	prior	MAP18 (2 views)
precision	0.99	-	-	0.99	0.7	0.98	0.85	0.8	0.99	0.79
recall(0.02)	0.48	0.52	0.42	0.59	0.67	0.51	0.66	0.71	0.49	**0.77**
recall(0.03)	0.61	0.64	0.48	0.73	0.80	0.64	0.78	0.84	0.66	**0.89**

Tabelle 1) Recall und Precision aller Schätzer (ML, prior und MAP). Es wurden jeweils die 5, 10, und 15 (alle) besten Armlets ausgewählt (ML5, ML10, etc.). Zusätzlich wurde das Modell um eine weitere Kameraperspektive und damit 3 weitere Posensignale erweitert (MAP18). ML5_le und ML5_ri beschreiben die Detektionsleistung für linke und rechte Körperhälfte getrennt.

während die Precision sich nicht signifikant verändert (Abfall um 0.01 auf 0.79). Die Hinzunahme weiterer Kameraperspektiven funktioniert also sehr gut, falls diese Perspektiven ebenfalls gute Posensignale erzeugen.

Verbesserung der Precision. Obwohl ein Recall von fast 90% für einen MAP Schätzer auf zwei Kameraperspektiven als zufriedenstellend bezeichnet werden kann, so zerstört eine Precision von 0.79 die Güte der Gesamtabschätzung, da diese zu viele Falschdetektionen enthält. Auch an dieser Stelle können zwei einfache, aber effiziente Heurisiken angewandt werden, um Falschdetektionen aus dem Signal zu entfernen. Zum Ersten wird eine Seitenbedingung für alle drei Schätzmethoden eingeführt, die nur dann die Position einer Schlüsselpose berechnen, wenn mindestens zwei verschiedenen Armlets die Vorhersage einer Schlüsselpose an derselben Stelle im Video machen würden, wobei natürliche eine kleine Abweichung erlaubt ist. Dies sortiert alle Schlüsselposenetektionen aus, die nur von einem einzigen Armlet vorgeschlagen werden. Zum Zweiten kann auf das finale Schlüsselposensignal die gleiche Heuristik angewendet werden wie schon bei den ursprünglichen Posensignalen. Dabei werden alle Detektionen aussortiert, die die Zugfrequenz des Schwimmers unnatürlich stark verändern. Beide Säuberungsmethoden zusammen vernichten fast alle Falschdetektionen, **sodass die Precision der besten MAP Schätzung schlussendlich auf einen akzeptablen Wert von 0.99 angehoben wird bei einem Recall von 0.9 für eine Abweichung von maximal 3% von der Grundwahrheit**. In dieser Konfiguration werden knapp 33% aller Schlüsselposen mit einer Abweichung von ±0 Halbbildern erkannt, weitere 45% der Detektionen weichen um ±1 Halbbild ab.

Abschließende Bemerkungen

Das vorgestellte System zur zeitlichen Bestimmung des Auftretens von Schlüsselposen ist selbstverständlich auch in der Lage, anfangs erwähnte kinematische Parameter zu bestimmen. Die Zugfrequenz des Schwimmers kann aus dem jeweils besten Posensignal erzeugt werden. Nach Erfahrung der Autoren gibt es für jeden Schwimmer zu jedem Zeitpunkt mindestens 4 Armlets, die die Zugfrequenz des Schwimmers (im Rahmen des menschlichen Fehlers) sehr gut abbilden. Abbildung

3 zeigt die Schwimmfrequenz eines Testschwimmers. Intrazyklische Frequenzen können über die Detektion verschiedener Schlüsselposen innerhalb eines Zyklus generiert werden.
Die abschließende Bewertung der intrazyklischen Frequenzen durch einen menschlichen Experten für Freistilschwimmer sowie die Ausweitung der Experimente für alle anderen Schwimmarten ist Gegenstand der weiteren Forschungen. Ebenso ist die Erweiterung des Ansatzes für andere Sportarten eine interessante Fragestellung für die Zukunft.

Danksagung

Wir bedanken uns beim IAT Leipzig, insbesondere bei der Fachgruppe Schwimmen, für die Zusammenarbeit und die Bereitstellung von Datenmaterial und fachspezifischen Informationen.
Dieses Forschungsprojekt wird durch das Bundesinstitut für Sportwissenschaft gefördert.

Gefördert durch:
Bundesinstitut für Sportwissenschaft
aufgrund eines Beschlusses des Deutschen Bundestages

Referenzen

[Zec12] Zecha, D., Greif, T., Lienhart, R.: Swimmer detection and pose estimation for continuous stroke-rate determination. In: Proc. SPIE. Volume 8304. (2012) 830410-830410-13

[Fis73] Fischler, M.A., Elschlager, R.A.: The representation and matching of pictorial structures. IEEE Trans. Comput. 22 (1973) 67-92

[Fel10] Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (2010) 1627-1645

[Wap95] Wladimir Wapnik: *The Nature of Statistical Learning Theory*, Springer Verlag, New York, NY, USA, 1995.

[Bou09] Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations. In: International Conference on Computer Vision (ICCV). (2009)

[Gki13] Gkioxari, G., Arbelaez, P., Bourdev, L., Malik, J.: Articulated pose estimation using discriminative armlet classifiers. 2013 IEEE Conference on Computer Vision and Pattern Recognition 0 (2013) 3342-3349

[Dal05] Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In Schmid, C., Soatto, S., Tomasi, C., eds.: International Conference on Computer Vision & Pattern Recognition. Volume 2., INRIA Rhône-Alpes, ZIRST-655, av. Del'Europe, Montbonnot-38334 (2005) 886-893

[Pro13] http://www.multimedia-computing.de/wiki/Swimmer_pose_estimation